EPID674 Epidemiologic Data Analysis using R

Descriptive Statistics, Tables

Kelly Bakulski, Lauren Middleton

Last compiled on November 02, 2022

## Descriptive Statistics, Tables, Misc.

# Install new packages

# Install packages. Do this only once.  
# We have this completed for you on Cloud, so you do not need to run this.  
options(repos="https://cran.rstudio.com" )  
install.packages("tidyverse")  
install.packages("here")  
install.packages("gtsummary")  
install.packages("flextable")  
  
# To avoid installing every time: change set up in curly brackets to eval=FALSE

# Set up: Query the current R environment, load relevant packages, load dataset created in Week 2

search() # list the packages, environments, or data frames

[1] ".GlobalEnv" "tools:quarto" "package:stats"   
 [4] "package:graphics" "package:grDevices" "package:utils"   
 [7] "package:datasets" "package:methods" "Autoloads"   
[10] "package:base"

ls() # list the objects

character(0)

##### Load these packages for the current session  
library(tidyverse)

── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──  
✔ ggplot2 3.3.6 ✔ purrr 0.3.5   
✔ tibble 3.1.8 ✔ dplyr 1.0.10  
✔ tidyr 1.2.1 ✔ stringr 1.4.1   
✔ readr 2.1.3 ✔ forcats 0.5.2   
── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
✖ dplyr::filter() masks stats::filter()  
✖ dplyr::lag() masks stats::lag()

library(here)

here() starts at /cloud/project

library(gtsummary)  
library(flextable)

Attaching package: 'flextable'  
  
The following objects are masked from 'package:gtsummary':  
  
 as\_flextable, continuous\_summary  
  
The following object is masked from 'package:purrr':  
  
 compose

sessionInfo() # record the versions of packages used in the current session

R version 4.2.1 (2022-06-23)  
Platform: x86\_64-pc-linux-gnu (64-bit)  
Running under: Ubuntu 20.04.5 LTS  
  
Matrix products: default  
BLAS: /usr/lib/x86\_64-linux-gnu/atlas/libblas.so.3.10.3  
LAPACK: /usr/lib/x86\_64-linux-gnu/atlas/liblapack.so.3.10.3  
  
locale:  
 [1] LC\_CTYPE=C.UTF-8 LC\_NUMERIC=C LC\_TIME=C.UTF-8   
 [4] LC\_COLLATE=C.UTF-8 LC\_MONETARY=C.UTF-8 LC\_MESSAGES=C.UTF-8   
 [7] LC\_PAPER=C.UTF-8 LC\_NAME=C LC\_ADDRESS=C   
[10] LC\_TELEPHONE=C LC\_MEASUREMENT=C.UTF-8 LC\_IDENTIFICATION=C   
  
attached base packages:  
[1] stats graphics grDevices utils datasets methods base   
  
other attached packages:  
 [1] flextable\_0.8.2 gtsummary\_1.6.2 here\_1.0.1 forcats\_0.5.2   
 [5] stringr\_1.4.1 dplyr\_1.0.10 purrr\_0.3.5 readr\_2.1.3   
 [9] tidyr\_1.2.1 tibble\_3.1.8 ggplot2\_3.3.6 tidyverse\_1.3.2  
  
loaded via a namespace (and not attached):  
 [1] Rcpp\_1.0.9 lubridate\_1.8.0 assertthat\_0.2.1   
 [4] rprojroot\_2.0.3 digest\_0.6.30 utf8\_1.2.2   
 [7] R6\_2.5.1 cellranger\_1.1.0 backports\_1.4.1   
[10] reprex\_2.0.2 evaluate\_0.17 httr\_1.4.4   
[13] pillar\_1.8.1 gdtools\_0.2.4 rlang\_1.0.6   
[16] uuid\_1.1-0 googlesheets4\_1.0.1 readxl\_1.4.1   
[19] rstudioapi\_0.14 data.table\_1.14.4 rmarkdown\_2.17   
[22] googledrive\_2.0.0 munsell\_0.5.0 broom\_1.0.1   
[25] compiler\_4.2.1 modelr\_0.1.9 xfun\_0.33   
[28] systemfonts\_1.0.4 base64enc\_0.1-3 pkgconfig\_2.0.3   
[31] htmltools\_0.5.3 tidyselect\_1.2.0 fansi\_1.0.3   
[34] crayon\_1.5.2 tzdb\_0.3.0 dbplyr\_2.2.1   
[37] withr\_2.5.0 grid\_4.2.1 jsonlite\_1.8.2   
[40] gtable\_0.3.1 lifecycle\_1.0.3 DBI\_1.1.3   
[43] magrittr\_2.0.3 scales\_1.2.1 zip\_2.2.1   
[46] cli\_3.4.1 stringi\_1.7.8 broom.helpers\_1.9.0  
[49] fs\_1.5.2 xml2\_1.3.3 ellipsis\_0.3.2   
[52] generics\_0.1.3 vctrs\_0.4.2 tools\_4.2.1   
[55] glue\_1.6.2 officer\_0.4.4 hms\_1.1.2   
[58] fastmap\_1.1.0 yaml\_2.3.6 colorspace\_2.0-3   
[61] gargle\_1.2.1 gt\_0.7.0 rvest\_1.0.3   
[64] knitr\_1.40 haven\_2.5.1

# Load saved NHANES dataset from session 02 on Data Management  
load(here(("nhanes\_class\_dataset.rda")), verbose = TRUE)

Loading objects:  
 nhanes

# Data description functions, numeric variables

# Base R method - quick, good for one variable at a time  
summary(nhanes$RIDAGEYR)

Min. 1st Qu. Median Mean 3rd Qu. Max.   
 0.00 11.00 31.00 34.33 58.00 80.00

sd(nhanes$RIDAGEYR)

[1] 25.50028

# Tidyverse method - more code, but can be used to get summary statistics on multiple variables  
nhanes %>%  
 summarise(age\_min = min(RIDAGEYR),  
 age\_25\_quart = quantile(RIDAGEYR, probs = 0.25),  
 age\_mean = mean(RIDAGEYR),  
 age\_median = median(RIDAGEYR),  
 age\_75\_quart = quantile(RIDAGEYR, probs = 0.75),  
 age\_iqr = IQR(RIDAGEYR),  
 age\_max = max(RIDAGEYR))

age\_min age\_25\_quart age\_mean age\_median age\_75\_quart age\_iqr age\_max  
1 0 11 34.33423 31 58 47 80

# Check your understanding!

Calculate descriptive statistics on the poverty income variable: INDFMPIR Specifically, calculate the minimum, 25th percentile, median, mean, 75th percentile, and maximum. \* Start by creating a new r code chunk \* Pick your method for calculating descriptive statistics!

# Calculate descriptive statistics on a categorical variable

# Base R method- quick, good for one variable at a time  
table(nhanes$race\_eth)

Non-Hispanic White Mexican American Non-Hispanic Black Other Hispanic   
 3150 1367 2115 820   
 Other Race   
 1802

# Tidyverse method - more code, but can be used to get summary statistics on combinations of multiple variables  
nhanes %>%   
 count(race\_eth) %>%  
 mutate(percent = n / sum(n) \* 100)

race\_eth n percent  
1 Non-Hispanic White 3150 34.039334  
2 Mexican American 1367 14.771990  
3 Non-Hispanic Black 2115 22.854982  
4 Other Hispanic 820 8.861033  
5 Other Race 1802 19.472660

# Reproducible, publication quality univariate tables

# Univariate table using gtsummary package  
colnames(nhanes)

[1] "SEQN" "RIASEX" "sex" "RIDAGEYR"   
 [5] "age\_groups" "RIDRETH1" "race\_eth" "INDFMPIR"   
 [9] "DMDEDUC3" "DMDEDUC2" "education\_youth" "education\_adult"  
[13] "education" "SDMVSTRA" "SDMVPSU" "LBXRBCSI"   
[17] "LBXWBCSI" "LBDLYMNO" "LBDNENO" "nlr"   
[21] "LBXIRN" "iron\_status" "URXUAS" "LBXBCD"   
[25] "LBXBPB" "LBXCOT" "cut\_groups"

nhanes %>%  
 select(RIDAGEYR,   
 sex,   
 race\_eth,  
 education,  
 LBXRBCSI,  
 LBXWBCSI,  
 LBDLYMNO,  
 LBDNENO,  
 nlr,  
 LBXIRN,  
 iron\_status,  
 URXUAS,  
 LBXBCD,  
 LBXBPB,  
 LBXCOT) %>% #drop SEQN and survey variables so we don't get summary statistics for the identifiers or weights  
 tbl\_summary(label = list(education ~ "Overall Educational Attainment",   
 race\_eth ~ "Race/ethnicity"),  
 statistic = all\_continuous() ~ "{mean} ({sd})")

Table printed with {flextable}, not {gt}. Learn why at  
https://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html  
To suppress this message, include `message = FALSE` in the code chunk header.

**Table** **:** Univariate table

| Characteristic | N = 9,2541 |
| --- | --- |
| Age in years at screening | 34 (26) |
| sex |  |
| Male | 4,557 (49%) |
| Female | 4,697 (51%) |
| Race/ethnicity |  |
| Non-Hispanic White | 3,150 (34%) |
| Mexican American | 1,367 (15%) |
| Non-Hispanic Black | 2,115 (23%) |
| Other Hispanic | 820 (8.9%) |
| Other Race | 1,802 (19%) |
| Overall Educational Attainment |  |
| Less than high school | 2,055 (26%) |
| Less than 5th grade | 1,142 (15%) |
| High school or GED | 1,480 (19%) |
| More than high school | 3,185 (41%) |
| Unknown | 1,392 |
| Red blood cell count (million cells/uL) | 4.73 (0.48) |
| Unknown | 1,726 |
| White blood cell count (1000 cells/uL) | 7.38 (5.11) |
| Unknown | 1,726 |
| Lymphocyte number (1000 cells/uL) | 2.50 (4.32) |
| Unknown | 1,731 |
| Segmented neutrophils num (1000 cell/uL) | 4.03 (1.73) |
| Unknown | 1,731 |
| Segmented neutrophils num (1000 cell/uL) | 1.88 (1.15) |
| Unknown | 1,731 |
| Iron frozen, Serum (ug/dL) | 86 (37) |
| Unknown | 3,332 |
| iron\_status |  |
| Deficient | 1,376 (23%) |
| Excessive | 142 (2.4%) |
| Normal | 4,404 (74%) |
| Unknown | 3,332 |
| Arsenic, Total - Urine (ug/L) | 16 (64) |
| Unknown | 6,462 |
| Blood cadmium (ug/L) | 0.37 (0.50) |
| Unknown | 1,741 |
| Blood lead (ug/dL) | 1.08 (1.29) |
| Unknown | 2,370 |
| Cotinine, Serum (ng/mL) | 40 (112) |
| Unknown | 2,153 |
| 1Mean (SD); n (%) | |

### Filter, order, bivariate descriptive statistics

# Ordering data

# Look at the current order of rows  
head(nhanes)

SEQN RIASEX sex RIDAGEYR age\_groups RIDRETH1 race\_eth INDFMPIR  
1 93703 2 Female 2 [0,16] 5 Other Race 5.00  
2 93704 1 Male 2 [0,16] 3 Non-Hispanic White 5.00  
3 93705 2 Female 66 (64,80] 4 Non-Hispanic Black 0.82  
4 93706 1 Male 18 (16,32] 5 Other Race NA  
5 93707 1 Male 13 [0,16] 5 Other Race 1.88  
6 93708 2 Female 66 (64,80] 5 Other Race 1.63  
 DMDEDUC3 DMDEDUC2 education\_youth education\_adult  
1 NA NA <NA> <NA>  
2 NA NA <NA> <NA>  
3 NA 2 <NA> Less than high school  
4 15 NA More than high school <NA>  
5 6 NA Less than high school <NA>  
6 NA 1 <NA> Less than high school  
 education SDMVSTRA SDMVPSU LBXRBCSI LBXWBCSI LBDLYMNO LBDNENO  
1 <NA> 145 2 NA NA NA NA  
2 <NA> 143 1 4.25 7.4 3.5 3.2  
3 Less than high school 145 2 5.48 8.6 3.4 4.2  
4 More than high school 134 2 5.24 6.1 1.5 3.7  
5 Less than high school 138 1 5.02 11.2 4.2 6.1  
6 Less than high school 138 2 4.59 6.0 1.9 3.6  
 nlr LBXIRN iron\_status URXUAS LBXBCD LBXBPB LBXCOT cut\_groups  
1 NA NA <NA> NA NA NA NA [0,8]  
2 0.9142857 NA <NA> NA 0.07 NA NA [0,8]  
3 1.2352941 92 Normal NA 0.24 2.98 0.028 (62,80]  
4 2.4666667 164 Normal NA 0.21 0.74 0.138 (8,20]  
5 1.4523810 91 Normal 5.09 0.14 0.39 0.555 (8,20]  
6 1.8947368 90 Normal 24.07 0.73 1.53 0.011 (62,80]

# Order by age and then sex  
nhanes %>%  
 arrange(RIDAGEYR, sex) %>%  
 head()

SEQN RIASEX sex RIDAGEYR age\_groups RIDRETH1 race\_eth INDFMPIR  
1 93748 1 Male 0 [0,16] 1 Mexican American 2.79  
2 93786 1 Male 0 [0,16] 4 Non-Hispanic Black 0.52  
3 93854 1 Male 0 [0,16] 3 Non-Hispanic White 3.67  
4 93865 1 Male 0 [0,16] 3 Non-Hispanic White 3.47  
5 93936 1 Male 0 [0,16] 3 Non-Hispanic White 3.74  
6 93962 1 Male 0 [0,16] 3 Non-Hispanic White 5.00  
 DMDEDUC3 DMDEDUC2 education\_youth education\_adult education SDMVSTRA SDMVPSU  
1 NA NA <NA> <NA> <NA> 142 2  
2 NA NA <NA> <NA> <NA> 147 1  
3 NA NA <NA> <NA> <NA> 143 1  
4 NA NA <NA> <NA> <NA> 135 1  
5 NA NA <NA> <NA> <NA> 145 1  
6 NA NA <NA> <NA> <NA> 145 1  
 LBXRBCSI LBXWBCSI LBDLYMNO LBDNENO nlr LBXIRN iron\_status URXUAS LBXBCD  
1 NA NA NA NA NA NA <NA> NA NA  
2 NA NA NA NA NA NA <NA> NA NA  
3 NA NA NA NA NA NA <NA> NA NA  
4 NA NA NA NA NA NA <NA> NA NA  
5 NA NA NA NA NA NA <NA> NA NA  
6 NA NA NA NA NA NA <NA> NA NA  
 LBXBPB LBXCOT cut\_groups  
1 NA NA [0,8]  
2 NA NA [0,8]  
3 NA NA [0,8]  
4 NA NA [0,8]  
5 NA NA [0,8]  
6 NA NA [0,8]

# Descending order by age and then sex  
nhanes %>%  
 arrange(desc(RIDAGEYR), sex) %>%  
 head()

SEQN RIASEX sex RIDAGEYR age\_groups RIDRETH1 race\_eth INDFMPIR  
1 93768 1 Male 80 (64,80] 3 Non-Hispanic White 3.04  
2 93869 1 Male 80 (64,80] 3 Non-Hispanic White 1.49  
3 93910 1 Male 80 (64,80] 3 Non-Hispanic White NA  
4 93930 1 Male 80 (64,80] 2 Other Hispanic NA  
5 93943 1 Male 80 (64,80] 3 Non-Hispanic White 5.00  
6 93963 1 Male 80 (64,80] 3 Non-Hispanic White 2.07  
 DMDEDUC3 DMDEDUC2 education\_youth education\_adult education  
1 NA 3 <NA> High school or GED High school or GED  
2 NA 5 <NA> More than high school More than high school  
3 NA 5 <NA> More than high school More than high school  
4 NA 1 <NA> Less than high school Less than high school  
5 NA 5 <NA> More than high school More than high school  
6 NA 2 <NA> Less than high school Less than high school  
 SDMVSTRA SDMVPSU LBXRBCSI LBXWBCSI LBDLYMNO LBDNENO nlr LBXIRN  
1 139 2 4.64 5.9 1.6 3.5 2.187500 64  
2 134 1 4.10 6.7 1.0 4.9 4.900000 105  
3 143 2 NA NA NA NA NA NA  
4 141 2 3.95 5.3 2.0 2.7 1.350000 123  
5 137 2 NA NA NA NA NA NA  
6 139 2 3.93 5.7 1.5 2.8 1.866667 90  
 iron\_status URXUAS LBXBCD LBXBPB LBXCOT cut\_groups  
1 Normal NA 0.12 1.65 0.011 (62,80]  
2 Normal 3.26 0.17 2.92 0.019 (62,80]  
3 <NA> NA NA NA NA (62,80]  
4 Normal 8.63 0.23 0.72 0.060 (62,80]  
5 <NA> NA NA NA NA (62,80]  
6 Normal NA 0.42 1.48 0.011 (62,80]

# Filter rows of the dataset

# Practice filtering: Keep only the female participants  
table(nhanes$sex) # What do you expect?

Male Female   
 4557 4697

nhanes\_subset <- nhanes %>%  
 filter(sex == "Female")  
dim(nhanes\_subset) # What do you get?

[1] 4697 27

# Keep only the participants who are iron deficient and older than 60 years using filter()  
table(nhanes$iron\_status, nhanes$RIDAGEYR >= 60) # What do you expect?

FALSE TRUE  
 Deficient 1000 376  
 Excessive 112 30  
 Normal 2915 1489

nhanes\_subset <- nhanes %>%  
 filter(iron\_status == "Deficient" & RIDAGEYR >= 60)  
dim(nhanes\_subset) # What do you get?

[1] 376 27

# Now can calculate univariate descriptive statistics on the subset datasets

# Calculate bivariate descriptive statistics

# Calculating statistics by groups - tidyverse  
# Note: good practice to add ungroup() after using it to avoid the dataset staying grouped  
nhanes %>%  
 group\_by(sex) %>%  
 summarise(age\_min = min(RIDAGEYR),  
 age\_25\_quart = quantile(RIDAGEYR, probs = 0.25),  
 age\_mean = mean(RIDAGEYR),  
 age\_median = median(RIDAGEYR),  
 age\_75\_quart = quantile(RIDAGEYR, probs = 0.75),  
 age\_iqr = IQR(RIDAGEYR),  
 age\_max = max(RIDAGEYR)) %>%  
 ungroup()

# A tibble: 2 × 8  
 sex age\_min age\_25\_quart age\_mean age\_median age\_75\_quart age\_iqr age\_max  
 <fct> <int> <dbl> <dbl> <int> <dbl> <dbl> <int>  
1 Male 0 10 34.1 30 58 48 80  
2 Female 0 11 34.5 32 57 46 80

# Calculate bivariate descriptive statistics on categorical variables

# Base R method  
table(nhanes$race\_eth, nhanes$sex)

Male Female  
 Non-Hispanic White 1583 1567  
 Mexican American 648 719  
 Non-Hispanic Black 1032 1083  
 Other Hispanic 400 420  
 Other Race 894 908

# Tidyverse method  
nhanes %>%   
 count(race\_eth, sex) %>%  
 mutate(percent = n / sum(n) \* 100)

race\_eth sex n percent  
1 Non-Hispanic White Male 1583 17.106116  
2 Non-Hispanic White Female 1567 16.933218  
3 Mexican American Male 648 7.002377  
4 Mexican American Female 719 7.769613  
5 Non-Hispanic Black Male 1032 11.151934  
6 Non-Hispanic Black Female 1083 11.703047  
7 Other Hispanic Male 400 4.322455  
8 Other Hispanic Female 420 4.538578  
9 Other Race Male 894 9.660687  
10 Other Race Female 908 9.811973

# Check your understanding!

Calculate the bivariate descriptive statistics between the iron status and age group variables: iron\_status and age\_groups \* Start by creating a new r code chunk \* Identify the types of variables to determine the descriptive statistics to calculate \* Try calculating them!

## Bivariate descriptive statistics, publication ready table

# Bivariate table by sex using gtsummary - will take about 5 seconds to run  
bivar\_nhanes <- nhanes %>%  
 select(RIDAGEYR,   
 sex,   
 race\_eth,  
 education,  
 INDFMPIR,  
 LBXRBCSI,  
 LBXWBCSI,  
 LBDLYMNO,  
 LBDNENO,  
 nlr,  
 LBXIRN,  
 iron\_status,  
 URXUAS,  
 LBXBCD,  
 LBXBPB,  
 LBXCOT) %>%  
 tbl\_summary(by = sex, #stratify by sex  
 label = list(RIDAGEYR ~ "Age (years)", #update the variable names  
 race\_eth ~ "Race/Ethnicity",  
 INDFMPIR ~ "Poverty-Income Ratio",  
 LBXRBCSI ~ "Red Blood Cell Count (million cells/uL)",  
 LBXIRN ~ "Iron (ug/dL)",  
 iron\_status ~ "Iron Status",  
 nlr ~ "Neutrophil:Lymphocyte Ratio"),  
 statistic = list(all\_continuous() ~ "{mean} ({sd})", #use mean and standard deviation for continuous variables  
 all\_categorical() ~ "{n} ({p}%)"), #use count and percentage for categorical variables  
 digits = list(all\_categorical() ~ c(0, 1), #adds no decimal places to counts and one decimal to percentages  
 all\_continuous() ~ 1), #adds one decimal place to mean and sd values  
 missing\_text = "Missing (n)"  
 ) %>%  
 add\_p() %>% #compares male vs female, does not include overall in p-value calculation  
 add\_overall() %>% #adds column with non-stratified summary statistics  
 modify\_header(label ~ "\*\*Variable\*\*") %>% #the asterisks bold the label  
 modify\_spanning\_header(c("stat\_1", "stat\_2") ~ "\*\*Sex\*\*") %>%  
 bold\_labels()  
  
#View table  
bivar\_nhanes

Table printed with {flextable}, not {gt}. Learn why at  
https://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html  
To suppress this message, include `message = FALSE` in the code chunk header.

**Table** **:** Bivariate table

|  | | Sex | |  |
| --- | --- | --- | --- | --- |
| Variable | Overall, N = 9,2541 | Male, N = 4,5571 | Female, N = 4,6971 | p-value2 |
| **Age (years)** | 34.3 (25.5) | 34.1 (25.8) | 34.5 (25.3) | 0.3 |
| **Race/Ethnicity** |  |  |  | 0.5 |
| Non-Hispanic White | 3,150 (34.0%) | 1,583 (34.7%) | 1,567 (33.4%) |  |
| Mexican American | 1,367 (14.8%) | 648 (14.2%) | 719 (15.3%) |  |
| Non-Hispanic Black | 2,115 (22.9%) | 1,032 (22.6%) | 1,083 (23.1%) |  |
| Other Hispanic | 820 (8.9%) | 400 (8.8%) | 420 (8.9%) |  |
| Other Race | 1,802 (19.5%) | 894 (19.6%) | 908 (19.3%) |  |
| **education** |  |  |  | 0.004 |
| Less than high school | 2,055 (26.1%) | 1,050 (27.3%) | 1,005 (25.0%) |  |
| Less than 5th grade | 1,142 (14.5%) | 570 (14.8%) | 572 (14.2%) |  |
| High school or GED | 1,480 (18.8%) | 745 (19.4%) | 735 (18.3%) |  |
| More than high school | 3,185 (40.5%) | 1,480 (38.5%) | 1,705 (42.4%) |  |
| Missing (n) | 1,392 | 712 | 680 |  |
| **Poverty-Income Ratio** | 2.4 (1.6) | 2.4 (1.6) | 2.4 (1.6) | 0.073 |
| Missing (n) | 1,231 | 618 | 613 |  |
| **Red Blood Cell Count (million cells/uL)** | 4.7 (0.5) | 4.9 (0.5) | 4.6 (0.4) | <0.001 |
| Missing (n) | 1,726 | 885 | 841 |  |
| **White blood cell count (1000 cells/uL)** | 7.4 (5.1) | 7.3 (2.5) | 7.5 (6.7) | 0.002 |
| Missing (n) | 1,726 | 885 | 841 |  |
| **Lymphocyte number (1000 cells/uL)** | 2.5 (4.3) | 2.4 (1.6) | 2.6 (5.8) | <0.001 |
| Missing (n) | 1,731 | 886 | 845 |  |
| **Segmented neutrophils num (1000 cell/uL)** | 4.0 (1.7) | 4.0 (1.7) | 4.1 (1.8) | <0.001 |
| Missing (n) | 1,731 | 886 | 845 |  |
| **Neutrophil:Lymphocyte Ratio** | 1.9 (1.2) | 1.9 (1.2) | 1.9 (1.1) | 0.3 |
| Missing (n) | 1,731 | 886 | 845 |  |
| **Iron (ug/dL)** | 86.3 (36.7) | 93.8 (35.9) | 79.1 (36.0) | <0.001 |
| Missing (n) | 3,332 | 1,683 | 1,649 |  |
| **Iron Status** |  |  |  | <0.001 |
| Deficient | 1,376 (23.2%) | 427 (14.9%) | 949 (31.1%) |  |
| Excessive | 142 (2.4%) | 91 (3.2%) | 51 (1.7%) |  |
| Normal | 4,404 (74.4%) | 2,356 (82.0%) | 2,048 (67.2%) |  |
| Missing (n) | 3,332 | 1,683 | 1,649 |  |
| **Arsenic, Total - Urine (ug/L)** | 16.2 (64.3) | 14.6 (32.0) | 17.8 (84.7) | <0.001 |
| Missing (n) | 6,462 | 3,180 | 3,282 |  |
| **Blood cadmium (ug/L)** | 0.4 (0.5) | 0.3 (0.4) | 0.4 (0.6) | <0.001 |
| Missing (n) | 1,741 | 891 | 850 |  |
| **Blood lead (ug/dL)** | 1.1 (1.3) | 1.3 (1.6) | 0.9 (0.9) | <0.001 |
| Missing (n) | 2,370 | 1,207 | 1,163 |  |
| **Cotinine, Serum (ng/mL)** | 40.4 (111.6) | 52.5 (128.0) | 28.9 (92.0) | <0.001 |
| Missing (n) | 2,153 | 1,102 | 1,051 |  |
| 1Mean (SD); n (%) | | | | |
| 2Wilcoxon rank sum test; Pearson's Chi-squared test | | | | |

# Save the table as a Word document - some formatting such as bold headers will be lost  
bivar\_nhanes %>%  
 as\_flex\_table() %>%  
 save\_as\_docx(path = here("summary\_stats\_table.docx")) #Export the table to Word using flextable package

# Missing data, date variables, for loops

# Do any of our variables have missing values?

# How many participants have complete data on iron measures?  
nhanes %>%  
 drop\_na(LBXIRN) %>% #exclude any participants with missing iron data  
 dim() #what are the dimensions after excluding missing iron data?

[1] 5922 27

# How many participants have complete data on all variables?  
nhanes %>%  
 drop\_na() %>% #exclude any participants with missing data  
 dim() #what are the dimensions after excluding missing data?

[1] 0 27

# What variables have the most missing data?  
summary(nhanes)

SEQN RIASEX sex RIDAGEYR age\_groups   
 Min. : 93703 Min. :1.000 Male :4557 Min. : 0.00 [0,16] :3250   
 1st Qu.: 96016 1st Qu.:1.000 Female:4697 1st Qu.:11.00 (16,32]:1519   
 Median : 98330 Median :2.000 Median :31.00 (32,48]:1349   
 Mean : 98330 Mean :1.508 Mean :34.33 (48,64]:1636   
 3rd Qu.:100643 3rd Qu.:2.000 3rd Qu.:58.00 (64,80]:1500   
 Max. :102956 Max. :2.000 Max. :80.00   
   
 RIDRETH1 race\_eth INDFMPIR DMDEDUC3   
 Min. :1.000 Non-Hispanic White:3150 Min. :0.000 Min. : 0.00   
 1st Qu.:3.000 Mexican American :1367 1st Qu.:1.040 1st Qu.: 3.00   
 Median :3.000 Non-Hispanic Black:2115 Median :1.920 Median : 6.00   
 Mean :3.234 Other Hispanic : 820 Mean :2.376 Mean : 6.35   
 3rd Qu.:4.000 Other Race :1802 3rd Qu.:3.690 3rd Qu.: 9.00   
 Max. :5.000 Max. :5.000 Max. :66.00   
 NA's :1231 NA's :6948   
 DMDEDUC2 education\_youth education\_adult   
 Min. :1.000 Length:9254 Length:9254   
 1st Qu.:3.000 Class :character Class :character   
 Median :4.000 Mode :character Mode :character   
 Mean :3.526   
 3rd Qu.:4.000   
 Max. :9.000   
 NA's :3685   
 education SDMVSTRA SDMVPSU LBXRBCSI   
 Less than high school:2055 Min. :134 Min. :1.000 Min. :2.320   
 Less than 5th grade :1142 1st Qu.:137 1st Qu.:1.000 1st Qu.:4.420   
 High school or GED :1480 Median :141 Median :2.000 Median :4.710   
 More than high school:3185 Mean :141 Mean :1.518 Mean :4.734   
 NA's :1392 3rd Qu.:145 3rd Qu.:2.000 3rd Qu.:5.030   
 Max. :148 Max. :2.000 Max. :7.840   
 NA's :1726   
 LBXWBCSI LBDLYMNO LBDNENO nlr   
 Min. : 1.900 Min. : 0.400 Min. : 0.400 Min. : 0.0933   
 1st Qu.: 5.800 1st Qu.: 1.800 1st Qu.: 2.800 1st Qu.: 1.1797   
 Median : 7.000 Median : 2.300 Median : 3.800 Median : 1.6667   
 Mean : 7.383 Mean : 2.502 Mean : 4.035 Mean : 1.8839   
 3rd Qu.: 8.500 3rd Qu.: 2.900 3rd Qu.: 4.900 3rd Qu.: 2.3077   
 Max. :400.000 Max. :358.800 Max. :35.200 Max. :21.5000   
 NA's :1726 NA's :1731 NA's :1731 NA's :1731   
 LBXIRN iron\_status URXUAS LBXBCD   
 Min. : 11.00 Deficient:1376 Min. : 0.16 Min. : 0.0700   
 1st Qu.: 61.00 Excessive: 142 1st Qu.: 3.10 1st Qu.: 0.1200   
 Median : 82.00 Normal :4404 Median : 6.09 Median : 0.2200   
 Mean : 86.25 NA's :3332 Mean : 16.25 Mean : 0.3736   
 3rd Qu.:106.00 3rd Qu.: 12.23 3rd Qu.: 0.4200   
 Max. :481.00 Max. :2802.29 Max. :13.0300   
 NA's :3332 NA's :6462 NA's :1741   
 LBXBPB LBXCOT cut\_groups   
 Min. : 0.050 Min. : 0.011 [0,8] :1904   
 1st Qu.: 0.460 1st Qu.: 0.011 (8,20] :1857   
 Median : 0.760 Median : 0.032 (20,42]:1868   
 Mean : 1.084 Mean : 40.370 (42,62]:1884   
 3rd Qu.: 1.300 3rd Qu.: 0.567 (62,80]:1741   
 Max. :42.480 Max. :1620.000   
 NA's :2370 NA's :2153

# Check function default settings for handling missing values

### How do functions handle missing values?  
  
#calculate summary statistics on iron  
summary(nhanes$LBXIRN)

Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
 11.00 61.00 82.00 86.25 106.00 481.00 3332

#calculate the mean of iron  
mean(nhanes$LBXIRN)

[1] NA

mean(nhanes$LBXIRN, na.rm = TRUE)

[1] 86.25464

#note the na.omit - remove all rows that include NAs  
mean(na.omit(nhanes$LBXIRN))

[1] 86.25464

#calculate a correlation between age and iron concentration  
cor(nhanes$RIDAGEYR, nhanes$LBXIRN)

[1] NA

cor(nhanes$RIDAGEYR, nhanes$LBXIRN, use="complete.obs")

[1] -0.01858113

# Try compiling results with a for loop

# Filter to the numeric variables, we'll be calculating numeric descriptive statistics  
nhanes\_numeric <- nhanes %>%  
 select\_if(is.numeric)  
  
# First prep the output dataset - set up blank columns  
out <- data.frame(matrix(nrow = ncol(nhanes\_numeric), ncol = 3))  
  
# Set the column names for the dataframe  
colnames(out) <- c("variable", "mean", "sd")  
  
# Pull the column names from nhanes and set them as the values for the first column  
out[, 1] <- colnames(nhanes\_numeric)  
  
# Then initiate the loop  
for (i in 1:ncol(nhanes\_numeric)) {  
 out[i, 2] <- round(mean(nhanes\_numeric[, i], na.rm = TRUE), digits = 2)  
 out[i, 3] <- round(sd(nhanes\_numeric[, i], na.rm = TRUE), digits = 2)  
}  
out

variable mean sd  
1 SEQN 98329.50 2671.54  
2 RIASEX 1.51 0.50  
3 RIDAGEYR 34.33 25.50  
4 RIDRETH1 3.23 1.28  
5 INDFMPIR 2.38 1.60  
6 DMDEDUC3 6.35 5.84  
7 DMDEDUC2 3.53 1.24  
8 SDMVSTRA 140.97 4.20  
9 SDMVPSU 1.52 0.50  
10 LBXRBCSI 4.73 0.48  
11 LBXWBCSI 7.38 5.11  
12 LBDLYMNO 2.50 4.32  
13 LBDNENO 4.03 1.73  
14 nlr 1.88 1.15  
15 LBXIRN 86.25 36.73  
16 URXUAS 16.24 64.32  
17 LBXBCD 0.37 0.50  
18 LBXBPB 1.08 1.29  
19 LBXCOT 40.37 111.60

# Date and Time objects

# Paste function

# create some dates to use in this exercise  
day <- c("1", "12", "13", "2")  
month <- c("1", "07", "08", "11")  
year <- c("1970","1980", "2000", "1959")  
  
# combine (paste) two or more variables that are parts of date  
birth\_days\_chr <- paste(month, day, year, sep = "/")  
str(birth\_days\_chr)

chr [1:4] "1/1/1970" "07/12/1980" "08/13/2000" "11/2/1959"

# Input date information as character, convert to date format

# convert to dates  
birth\_days\_date <- as.Date(birth\_days\_chr, format = "%m/%d/%Y")  
str(birth\_days\_date)

Date[1:4], format: "1970-01-01" "1980-07-12" "2000-08-13" "1959-11-02"

as.numeric(birth\_days\_date)

[1] 0 3845 11182 -3713

birth\_days\_month <- format(birth\_days\_date, "%b %d, %Y")  
birth\_days\_month

[1] "Jan 01, 1970" "Jul 12, 1980" "Aug 13, 2000" "Nov 02, 1959"

# day of the week  
format(birth\_days\_date, "%a-%d%b%y")

[1] "Thu-01Jan70" "Sat-12Jul80" "Sun-13Aug00" "Mon-02Nov59"

weekdays(birth\_days\_date)

[1] "Thursday" "Saturday" "Sunday" "Monday"

# Calculate age as of today’s date

date\_today <- Sys.Date() # Sys.time or Sys.Date: Current time/Date  
date\_today

[1] "2022-11-02"

current\_age <- (date\_today - birth\_days\_date) / 365.25  
current\_age

Time differences in days  
[1] 52.83504 42.30801 22.22040 63.00068

current\_age <- trunc(as.numeric(current\_age))  
  
  
# create data frame  
bd <- data.frame(Character = birth\_days\_chr, Standard\_Date = birth\_days\_date, Numeric\_Date = as.numeric(birth\_days\_date), Age = current\_age)  
bd

Character Standard\_Date Numeric\_Date Age  
1 1/1/1970 1970-01-01 0 52  
2 07/12/1980 1980-07-12 3845 42  
3 08/13/2000 2000-08-13 11182 22  
4 11/2/1959 1959-11-02 -3713 63

# Remember to save your R script!

# To exit R

# q()  
## if you close R, you will be asked to save your workspace image